Spread, comparison of batches, stragglers

		Page
3A.	Spread and its indicators	3-1
3B.	Plotting batches for comparison	3-9
3C.	Effective choice of expression	3-14
3D.	Tippett's warp breaks	3-21
3E.	A rough look for stragglers	3-30
3F.	Typical squares and mean squares	3-35
3G.	How far have we come?	3-40

*****

3K.	How to choose the next trial expression using spread	3-42
3L.	How to choose the next trial expression using symmetry (if you must)	3-45

*****

3P.	IEDA processor	

*****

3X.	Another example	3-51
3Y.	Problems	

Location of tables and plots for 3

Number	Page
1	3-3
2	3-6
3	3-7
4	3-8
5	3-10
6	3-12
7	3-13
8	3-16
9	3-18
10	3-19
11	3-20
12	3-23
13	3-24
14	3-25
15	3-27
16	3-28
17	3-29
18	3-31
19	3-34
20	3-44
21	3-47
22	3-49
23	3-52
24	3-54
25	3-55
26	3-56
27	3-57

3

Spread, comparison of batches, stragglers

3A. Spread and its indicators

Given data, our first attention goes -- as we have seen -- to "How big are the values? Roughly, where does it fall?". Typical values help us to answer this numerically; simple plots of display and relationship help us to answer this visually. Both also help to generate residuals -- indications of how deviant, both in direction and amount, individual pieces of data appear to be. We have begun what will be a steady emphasis on the importance of residuals and the need to choose typical values so that, as far forth as is reasonable, individual residuals will tell us about individual pieces of data.

Once we have decided to try to separate typical value and residual as best we can, there is a place for summaries of the behavior of residuals. In most cases, a summary of residuals which keeps their signs ought not to be needed. The typical value of a batch of residuals should be much smaller than the residuals themselves. After all, the residuals were figured as + and - deviations from a typical value. Rather, what we need to summarize first is how widespread or tightly packed are the residuals.

The simplest case is when we have a single batch of observations, which we naturally refer to a single typical

value. Here any measure of collective residual size has to be close to being a measure of how spread out are the values in the batch. So we can reasonably begin by asking about spread, planning to return to ask about other measures of size of residuals.

Let us return to the 10 bowling scores. Table 1 shows them displayed in a way that makes it easy to identify several simple and useful measures of spread:

- the "range" -- or range of ends -- is the difference between the end values.
- the "next range" -- or range of nexts -- is the difference between the values next to each end.
- the "easy midspread" -- or range of hinges -- is the difference between the hinges (recall that if any values appear partially in the exact midmean, they are the hinges, otherwise the hinges are halfway between the closest values not in the body and the most extreme values in the body.
- the "exact midspread" -- or range of quarters -- sometimes called the interquartile distance, is the difference between the two quarter points, or quartiles, whose calculation we will need to explain.

These simple measures of spread fall naturally into two pairs. One pair looks at the edges of the batch, and is sensitive to relatively extreme deviations. Its members are:

- the range.
- the next range.

table 1 of chapter 3

The 10 Bowling Scores Displayed and Some Indicators of Spread

(1) The end values

(2) The next-to-end values

(3) Other values in the wings (none for so small a sample)

(4) The hinges

(5) The body

(6) The quarter points or quartiles

	(1)	(2)	(3)	(4)	(5)	(6)
	214					
		210	---			
				207		$207.75 = \frac{1}{4}(210) + \frac{3}{4}(207)$*
					203	
					198	
					193	
					192	
				191		$189.5 = \frac{1}{4}(185) + \frac{3}{4}(191)$*
		185	---			
	184					
Diffs	30 (Range)	25 (Next range)	(---)	16 (Easy midspread)	(---)	18.25** (Exact midspread)

* See text or next table for explanation.

** Also can be found as $\frac{3}{4}(16) + \frac{1}{4}(25)$, where 25 is the range of the two observations just outside the hinges and $\frac{3}{4}$ comes from $\frac{11}{4} = 2$ and $\frac{3}{4}$.

The other pair looks only part way out on each side of the batch, avoiding, among others, any few very extreme deviations there may be. Its members are:

- the easy midspread.
- the exact midspread.

One or two very extreme values pose a problem. Sometimes it is important to notice that they are there -- sometimes it is important not to be distracted by them. These two situations are both important; we need both indicators of spread that seize on extreme values and indicators that avoid them. But why two of each?

A single very extreme deviation occurs in more batches than we might think -- and for a wide variety of reasons. The next range dodges the effects of such a single "way out" value, while the range responds to it. Sometimes we want one, sometimes the other.

Among indicators that look at intermediate values, the reason for two choices is quite different. The easy midspread is like the easy midmean. It is easy to find and very useful in analysis, especially hand analysis. The exact midspread is like the exact midmean. It is more precise, more publishable, takes more effort, and might as well be computed if a machine is to do the work.

The way in which the quarter points -- and thus the exact midspread -- are computed in table 1 should seem mysterious. Where do the 1/4 and 3/4 come from? The idea, however, is simple:

- the 10 observations divide all possible values into 11 intervals. To find quarter points we should count in 1/4 of 11, that is 2.75, intervals. Table 2 shows this more graphically.

If we repeat the display of table 1 for the 15 weights of "nitrogen" found by Rayleigh (see problem 6 at page 1-56), as is done in table 3 we can see a number of things:

- the quarter points do not always require interpolation.
- the quarter points and hinges are sometimes the same (when the sample size is 3, 7, 11, 15, 19, 23, ...).
- just looking at the values of a measure of spread need not tell us all that is going on. We know that Rayleigh's weights fall into seven higher values (from air) and eight lower values (from other sources of nitrogen). No one measure of spread tells us this in any direct way. (Of course, if we knew how much spread was likely to come from experimental technique, the size of any of them would warn us that something is going on.)

Looking at the two batches into which Rayleigh's weights are naturally divided, gives us some idea of how the four measures of spread behave -- even coalesce -- in slightly smaller batches. This is done in table 4, where we see:

- no entries in column (3)
- for a batch of seven, the appearance of the same entries in columns (2), (4), and (6). (This also happens

table 2 of chapter 3

10 Values Make 11 Intervals -- The Quarter Points of the Bowling Scores

(1) The 10 values

(2) The 11 intervals

(3) The inside ends of two intervals at each end

(4) The inside ends of three intervals at each end

(5) The inside ends of 1/4 (11) = 2.75 intervals at each end

(1)	(2)	(3)	(4)	(5)
	←			
214				
	←			
210		210		
	←			$207.75 = \frac{1}{4}(210) + \frac{3}{4}(207)$
207			207	
	←			
203				
	←			
198				
	←			
193				
	←			
192				
	←			
191			191	
	←			$189 = \frac{1}{4}(185) + \frac{3}{4}(191)$
185		185		
	←			
184				
	←			

table 3 of chapter 3

Rayleigh's 15 Weights of "Nitrogen" and
the Four Measures of Spread

(1) The end values

(2) The next-to-end values

(3) Other wing values

(4) The hinges

(5) Body values

(6) The quarter points

(1)	(2)	(3)	(4)	(5)	(6)
2.31030					
	2.31028				
		2.31024			
			2.31017		2.31017
				2.31010	
				2.31001	
				2.30986	
				2.30182	
				2.30143	
				2.29940	
				2.29890	
			2.29889		2.29889
		2.29889			
	2.29849				
2.29816					
.01214 (Range)	.01179 (Next range)		.01128 (Easy midspread)		.01126 (Exact midspread)

table 4 of chapter 3

Rayleigh's Batches of Seven and Eight Weights of Nitrogen and the Four Measures of Spread

	(1) End	(2) Next	(3) Wing	(4) Hinge	(5) Body	(6) Quarter
(a) THE BATCH OF 7 FROM AIR						
	2.31030					
		2.31028	---	2.31028		2.31028
					2.31024	
					2.31017	
					2.31010	
		2.31001	---	2.31001		2.31001
	2.30986					
Range:	.00044	.00027		.00027		.00027
(b) THE BATCH OF 8 FROM OTHER SOURCES						
	2.30182					
		2.30143	---	(2.30042)		2.30092
					2.29940	
					2.29890	
					2.29889	
					2.29889	
		2.29849	---	(2.29869)		2.29859
	2.29816					
Range:	.00366	.00294		.00173		.00233**

**Also can be found as $\frac{1}{4}$ (.00051) + $\frac{3}{4}$ (.00294), where .00294 is the range of the values next outside the hinges and .00051 is the range of the value next inside the hinges. This works in samples of size 4, 8, 12, 16, 20, ...

for batches of 6 or 5. For batches of 4, the next range is barely defined.)

- for a batch of eight, the appearance of hinges that are not values but rather midpoints between values. (This also happens for batches of 4, 12, 16, 20, ...)

3B. Plotting batches for comparison

How are we to show these indications of spread, or most of them, in a visual display? How much can we try to show without confusion? When do we want to see the original data, when a schematic summary? All these questions are natural and important -- the simple examples we have been considering can show us much about the answers to all of them.

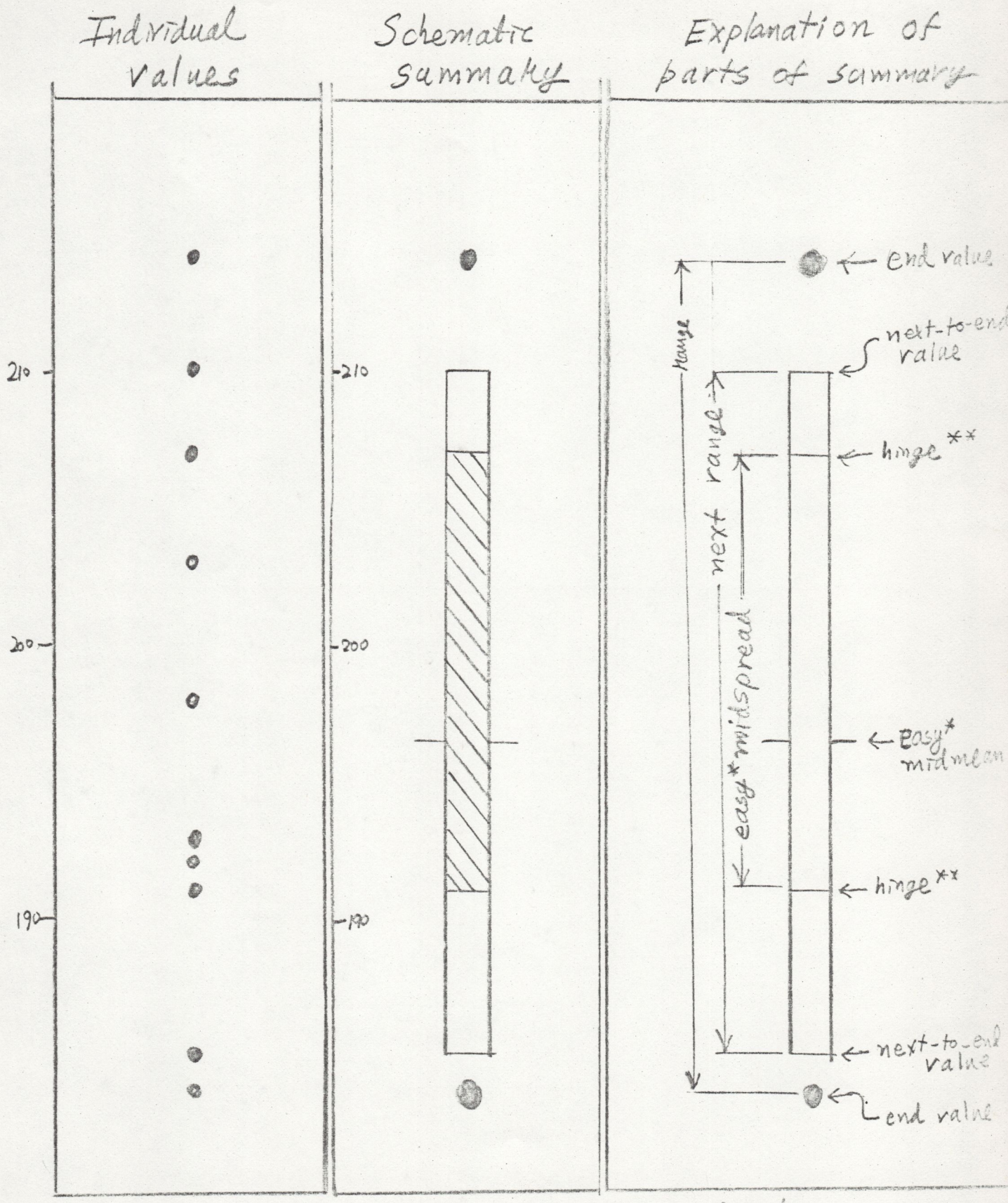
Plot 5 shows the 10 bowling scores in two versions. The left-hand panel shows the individual values, while the center panel, as explained to the right, shows:

- the midmean.
- the hinges -- and thus the easy midspread.
- the next-to-the end values -- and thus the next range.
- the end values -- and thus the range.

In this instance, the schematic summary tells us essentially all that the individual values do, doing this in a more eye-catching way -- a way easier to grasp at a glance.

Notice that the descriptive panel allows for either an "easy" version or a "publication" version, one with:

The 10 Bowling Scores displayed



* or exact

** or quarter point

- easy midmean,
- hinges,
- easy midspread,

the other with:

- exact midmean
- quarter points
- exact midspread.

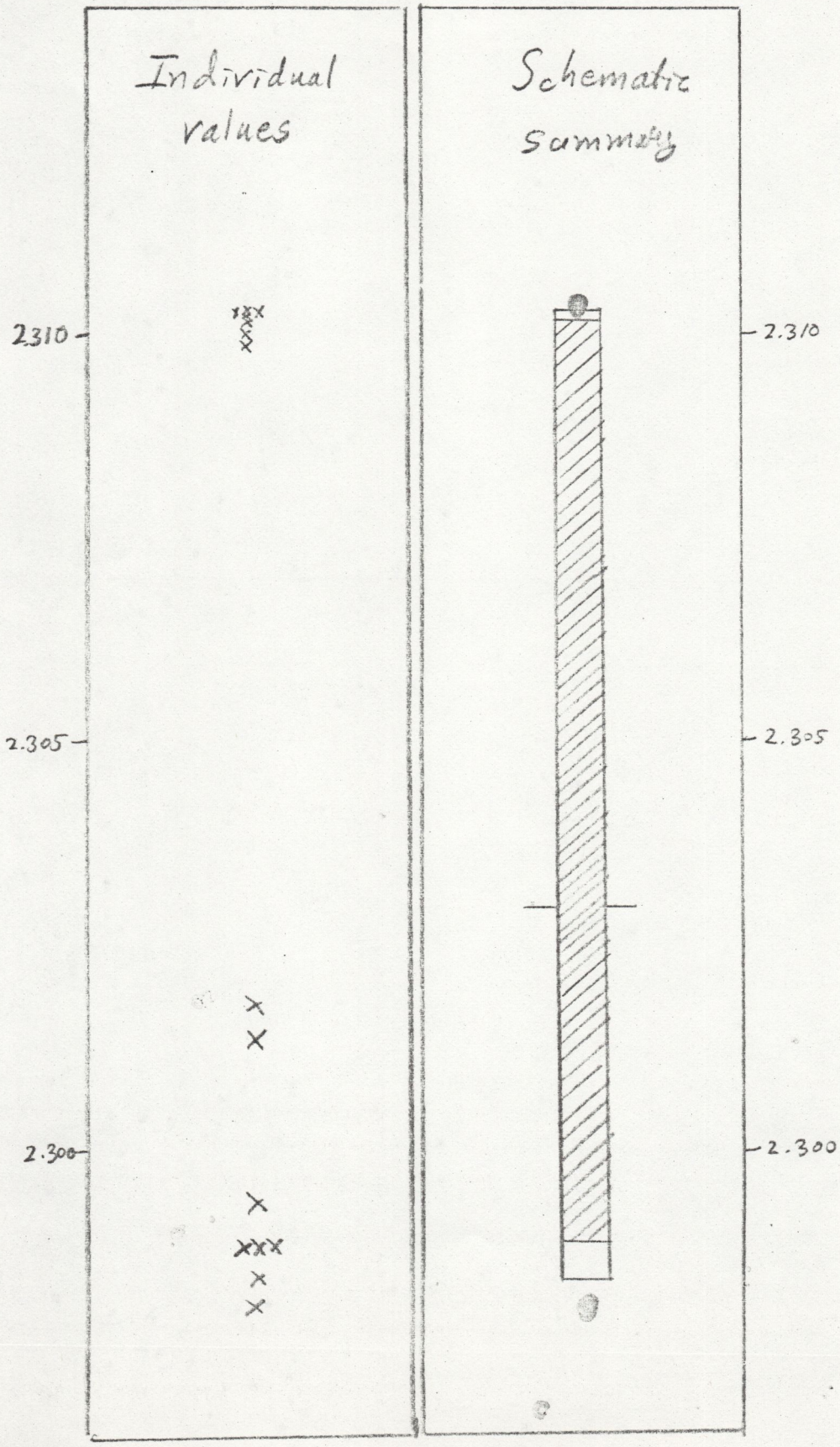
Plot 6 shows the same comparison for Rayleigh's 15 weights of nitrogen, treated as a single batch. Here the main issue is the violent separation into two quite isolated subgroups. This is made quite clear by the individual values -- and almost completely covered up by the schematic summary. (Only almost, because the experienced viewer -- finding end, next, and hinge values so close together -- is likely to become suspicious that he should see more detail.) Clearly we cannot rely on schematic summaries to call our attention to separations well within the batch, we must use some other approach. (Just looking at residuals does quite well.)

Plot 7 uses the schematic summaries for one of the purposes for which they are best fitted: comparison of two or more batches. In it the two batches of Rayleigh's weights, one for "nitrogen" from air, the other for "nitrogen" from other sources, are set out and compared. Clear to every eye are:

- the relatively huge difference between the batches.
- the tendency of the "other sources" weights to straggle upward -- noticeable but far from offering any basis for explaining the air values.

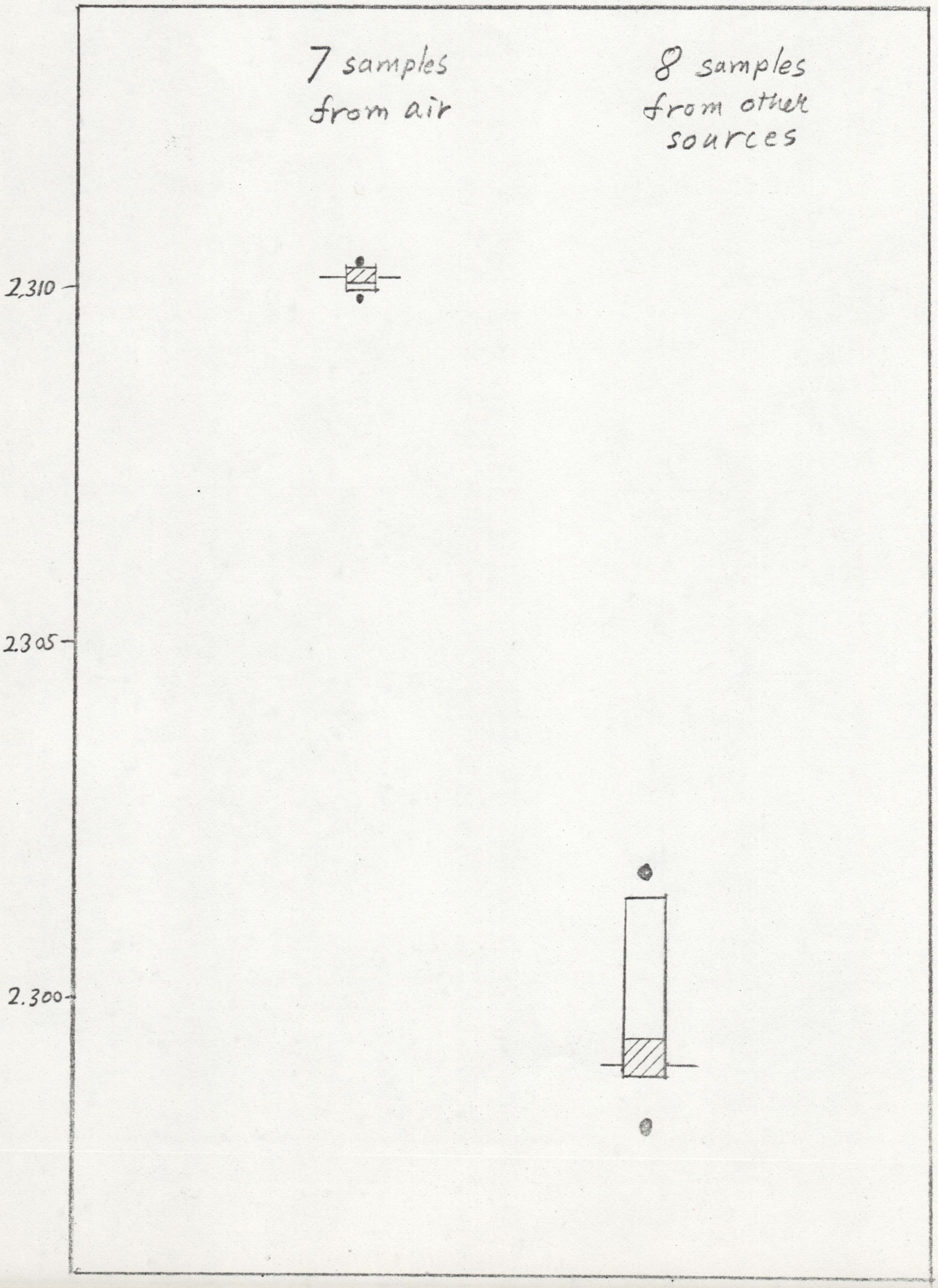
plot 6 of chapter 3

Rayleigh's 15 weights of "Nitrogen"



plot 7 of chapter 3

Rayleigh's two batches of weights of "nitrogen"



- a possible greater variability of the "other sources" weights, over and beyond this straggling.

Here the schematic summaries have done everything that we could ask.

Clearly we can use the same approach in comparing 3, 4, or any reasonable number of batches. (When the number of batches becomes large, we will probably want to squeeze our filled and open bars down to solid and dotted lines.) As a visual approach to most comparisons, especially when supplemented by more detailed looks at the residuals, these schematic summaries work very well.

3C. Effective choice of expression

In the last chapter we learned something of the great gains in visual impact that can come from more careful choice of what values are to be plotted. There, since we were dealing with plots of relationship, our emphasis was on straightness and flatness. Sometimes plots of comparison are also plots of relationship -- for instance when we have repeated measurements of some physical property at or near several temperatures. In such instances, all the considerations of chapter 2 apply, and dominate all others. The first thing we do is to choose plotting scales to make things nearly straight and then to form differences (residuals) to make them nearly flat. Only after we have gone as far as we can in such a direction should we listen to other considerations.

There are many instances, however, where no issue of relationship arises, where we want to bring eye and brain to

as clear and clean a comparison as we know how to make. Here we should pay attention to other considerations from the beginning.

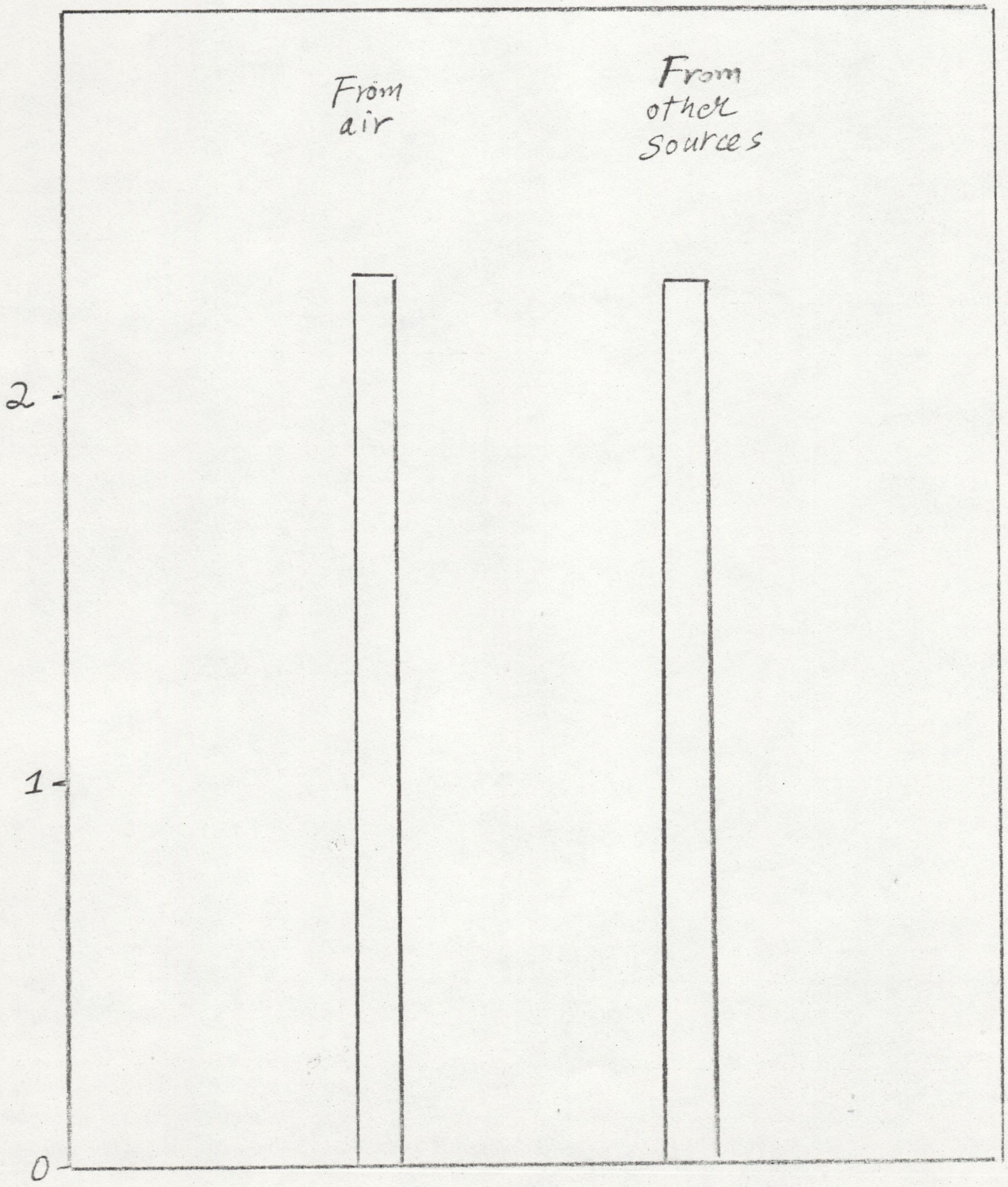
As often happens, detective work and newspaper reading are not the same thing here. Rayleigh's weights are a case in point. Many books on graphical presentation stress getting the proper impression of relative size across to the reader. This is quite sound for most articles in the daily newspaper, where making small differences more visible by starting the vertical scale far from zero is usually thought of as a "journalistic" or "political" trick. We need to remember that doing just this is also a scientific tool. If we had started the vertical scale at zero, and used bar graphs, Rayleigh's results -- so clear in plot 7 -- would have appeared as in plot 8. Whatever plot 8 may look like, it is neither a new chemical element nor a Nobel prize.

From the point of view of freight charges for compressed gases, plot 8's impression of "negligible difference" gives a useful impression, but it takes plot 7 to make clear the scientific issue. There is often no substitute for the detective's microscope.

When comparison involves no question of relationship so that we are free to use other grounds as a basis for choosing expression, we will usually clear up our perceptions when we make the batches to be compared more nearly equal in spread. There may be no way to do this, it may be possible but difficult, or it may be easy. When we can do it, we usually gain by it.

plot 8 of chapter 3

Rayleigh's weights again



One way to do it is to choose the way our data is expressed, to reform its expression. Table 9 gives a simple example from Winsor and Clarke's (1940) work on the catch of plankton using different kinds of nets. Plot 10 shows the direct comparison, which is not too neat. The difference in typical catch of the three kinds of plankton is all too obvious, but the great differences in spread tend to mask any detailed examination.

Especially since the natural change in catch is by such-and-such %, not by so-and-so many plankton, it seems natural to try the use of logs. When we do, we find the results in plot 11. Now our spreads are more nearly constant, and we get a fairly detailed idea of how the catches of the three kinds of plankton compare. In particular, apparent tendencies toward:

- one high and one low value for Kind IV,
- two high and two low values for Kind III,

are reasonably evident, as is the rather clearer resolution of the difference between Kind I and Kind IV, in comparison with the difference between Kind IV and Kind III.

Plot 11 shows some tendency for the spread to decrease as the typical value, also on the log scale, increases. Were this tendency as pronounced as the opposite tendency, shown in plot 9, of spread to increase as the typical number increases, we might be tempted to "split the difference" and try the square roots of the estimated counts with a reasonable anticipation of making the spreads more nearly constant.

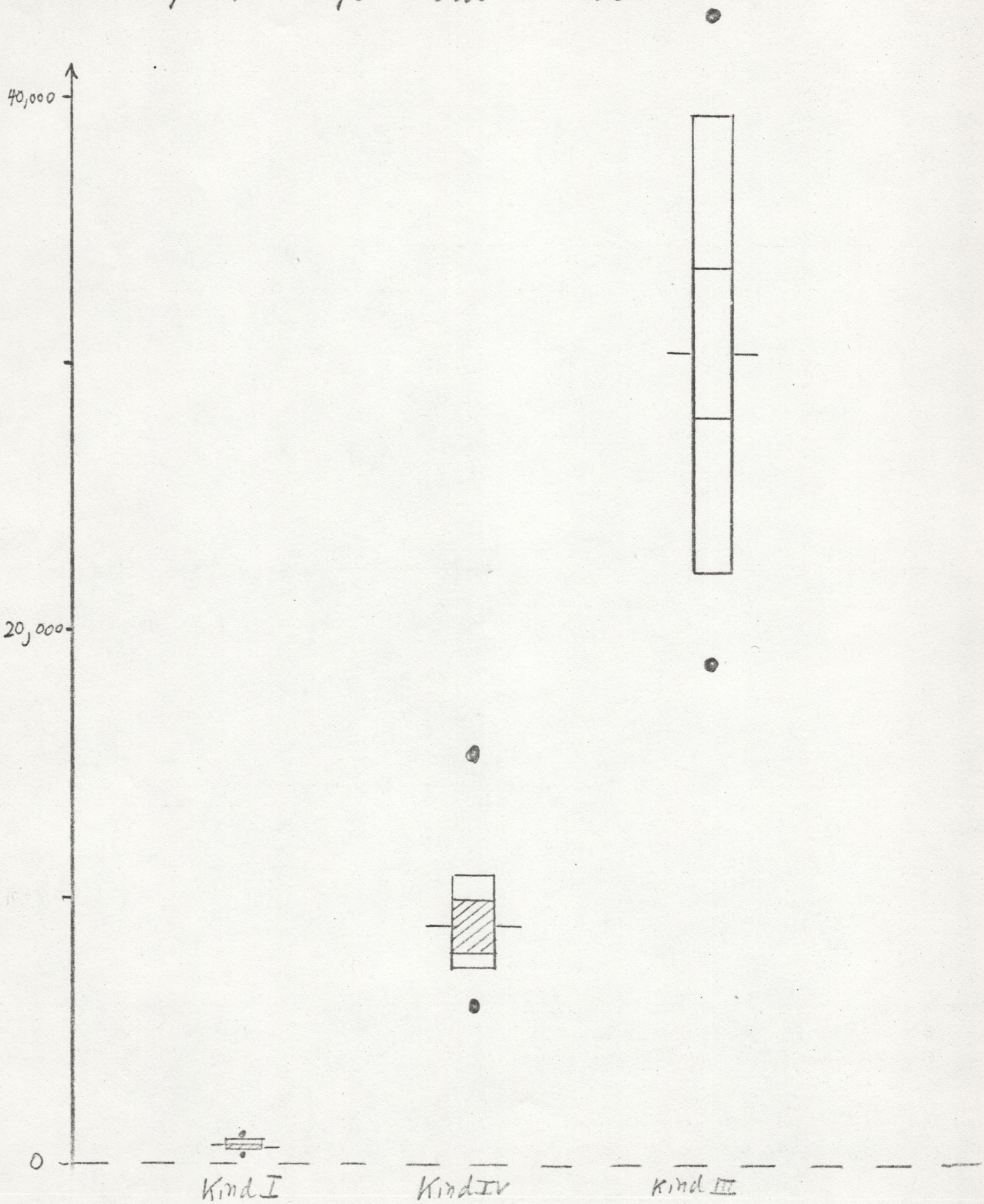
table 9 of chapter 3

Estimated Number of Three Kinds of Plankton Caught in Six Hauls of Each of Two Nets

Kind I	Kind IV	Kind III
1,050	15,500	43,300
1,020	11,000	39,500
895	10,200	34,600
845	9,830	32,800
760	9,650	32,800
620	9,250	31,400
540	8,900	29,000
537	8,600	28,800
497	8,260	28,100
470	7,900	27,800
428	7,600	22,300
387	6,060	18,900

From: C. P. Winsor and G. L. Clarke 1940. Journal of Marine Research (Sears Foundation) 3: 1. See also: G. W. Snedecor 1946. Statistical Methods, 4th ed., page 451.

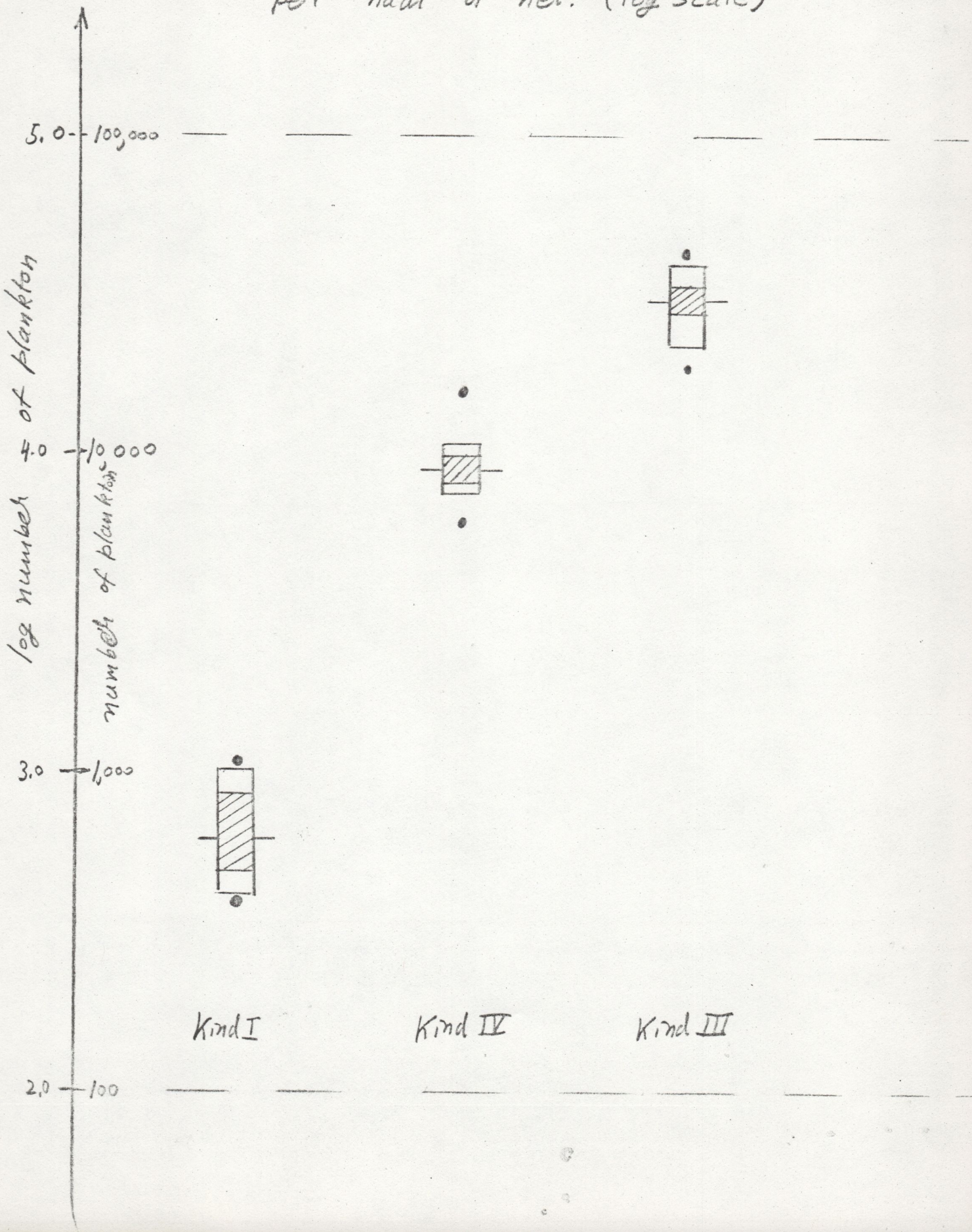
Winsor and Clarke's estimated number of plankton per haul of net



Source: See table 8

plot **11** of chapter 3

Winsor and Clarke's estimated number of plankton per haul of net. (log scale)



As it is, the square root would probably give more equal spreads than the log, but the cube root should make the spreads even more constant. For the present data, the need to make the spreads more equal than in plot 10 is not pressing, while logs have certain advantages we shall see when we come to later appearances of this data. Accordingly, logs seem the proper choice.

As this example has illustrated, proper choice of expression can be a great help in comparing batches of values. When relationship is not involved -- or when its demands have been met -- we can usefully be guided by trying to keep spread from changing systematically with typical value.

((**Occasionally we want to compare batches whose typical values are so much alike as to make constancy of spread about the same for a wide variety of expressions. Often this means that choice of expression is not important. Sometimes it means that we can be usefully guided by trying to make the batches look symmetrically spread out. Symmetry of pattern is, however, much less useful, much weaker, and comes much later in priority than does constancy of spread. Both, of course, give place to straightness and flatness of relationship.))

3D. Tippett's warp breaks

It is time to look at a somewhat larger, more nearly life-sized example. Tippett [1950, p. 106] gave nine observations on the number of warp breaks during a fixed amount of weaving for each of six types of warp (AL,AM,AH,BL,BM,BH), differing

in kind of cotton and tightness of twist. We shall take these 54 counts, listed in table 12, as a basis for illustrating a variety of visual approaches -- most of which will involve some arithmetic.

Plot 13 shows the 54 values sorted out into the six batches that Tippett labeled AL to BH. This plot shows us quite a lot -- too much for many purposes. We are given ideas about differences in level among the six batches; AL rather clearly runs high, while BH runs low. We are given ideas about the spreads of the various batches of nine values, AL appearing broader and BH narrower. We are even given ideas about the internal pattern of the observations. And there may even be clues about other sorts of behavior.

Focusing our attention a little more closely is very often helpful. Plot 14 shows the schematic summaries for the six batches, arranged in order of increasing midmeans. We see from this, perhaps more clearly, many of the same things we could have seen in plot 12:

- spread does seem to increase with midmean.
- body lengths are rather variable, perhaps no more than might have been expected from the size of the batch.
- one next-to-end value seems "far out" (the high one for BL).
- a few end values appear to be "far out", all on the high side.

Thus, both change of spread with midmean and apparent lack of symmetry point toward a change of expression.

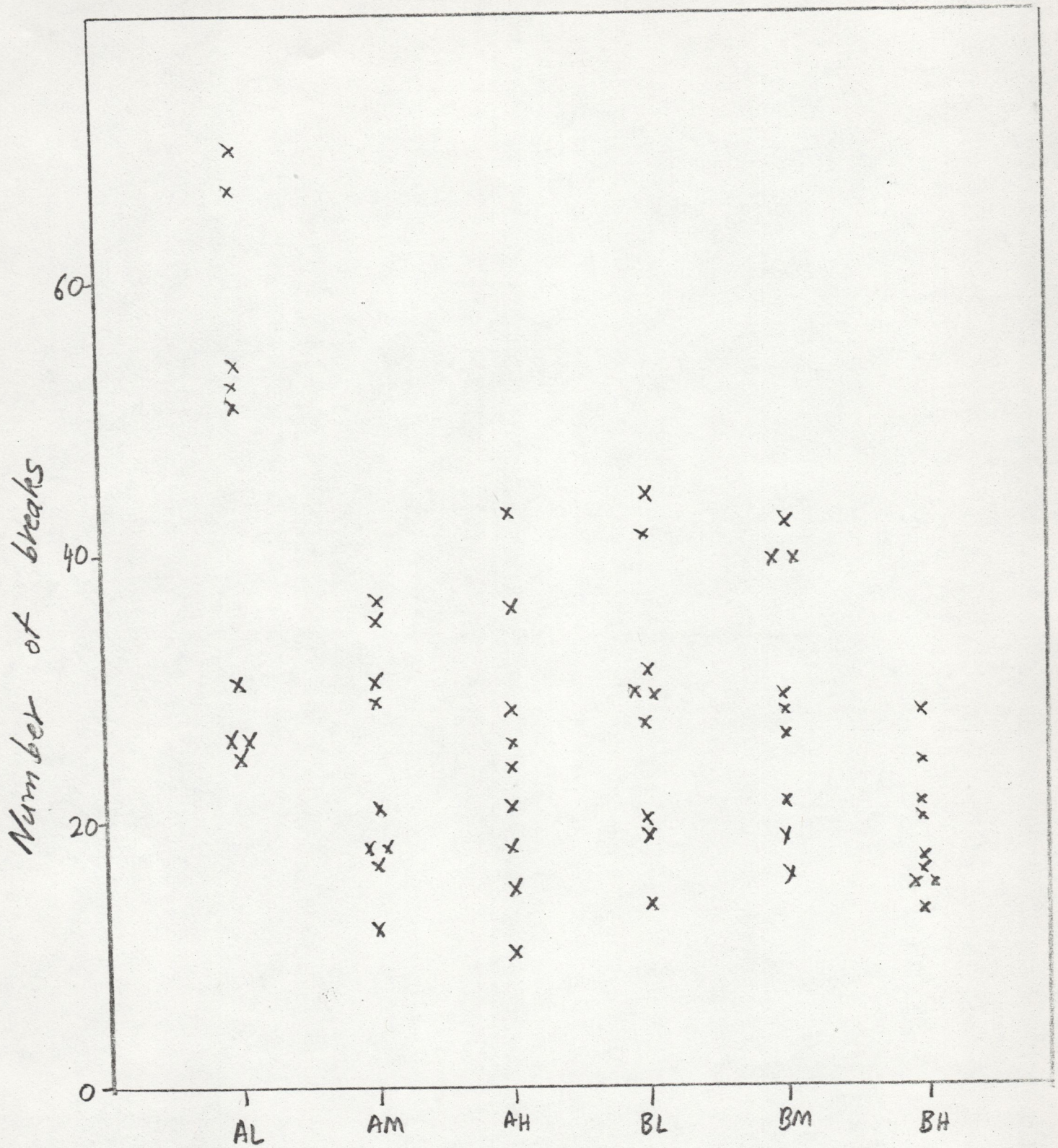
table 12 of chapter 3

The 54 Counts of Warp Breaks

AL	AM	AH	BL	BM	BH
26	18	36	27	42	20
30	21	21	14	26	21
54	29	24	29	19	24
25	17	18	19	16	17
70	12	10	29	39	13
52	18	43	31	28	15
51	35	28	41	21	15
26	30	15	20	39	16
67	36	26	44	29	28

Source: Page 106 of: L. H. C. Tippett 1950. Technological Applications of Statistics. New York: John Wiley and Sons. London: Williams and Norgate.

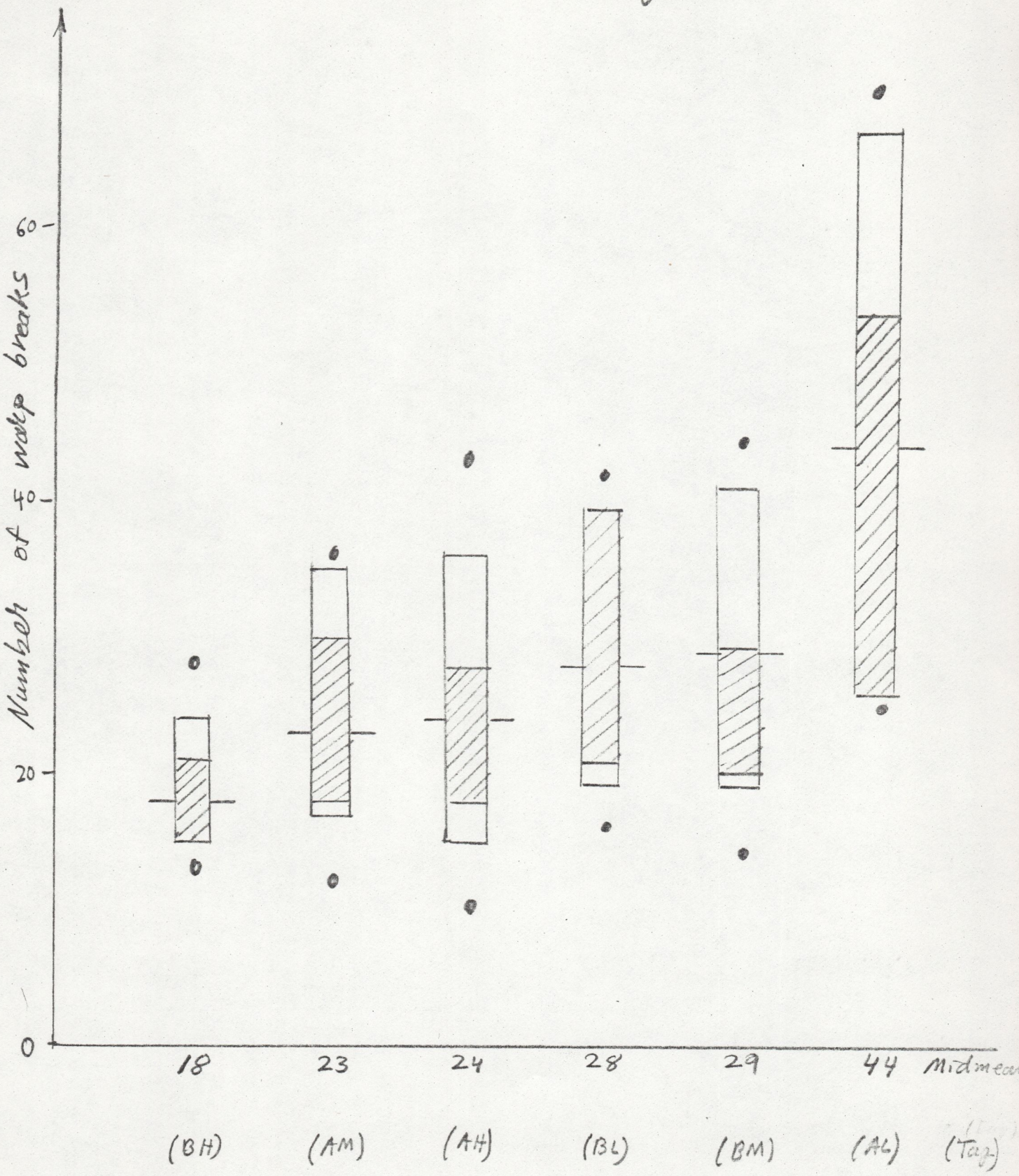
plot 13 of chapter 3



Tippett's data on number of warp breaks

plot 14 of chapter 3

Tippett's warp breaks again



Let us believe that spread does increase with midmean when we use warp counts. Accordingly, let us try:

- square roots of counts as the manner of expression.
- plotting the deviations of these square roots from their midmeans as a way to focus our attention on what happens within batches.
- using the "publication" version of the schematic summary.

Table **15** gives the numbers, plot **16** gives the picture. Rather clearly there is still an apparent tendency for spread to increase with midmean. To get a better idea of the strength of this apparent tendency, it will help to do more than merely order these schematic summaries.

We can learn more by plotting them against the value of the midmeans. (In this case, starting the picture from midmean = 0 IS worthwhile.) When we attempt this, a glance at the six midmeans:

4.22, 4.77, 4.83, 5.32, 5.26, 6.46

indicates that two pairs are so close together that we will clarify the picture by combining each pair (by taking means). This has been done in the two right-most columns of table 15, so that we have all the numbers required for plot **17**. Everything in this plot seems to be converging toward the origin, so that we feel that the appearance is very definite and quite strong. Presumably we should pass on to trying, as the next step stronger than square roots, logs of counts. This we shall leave to the reader. (See problems ---------- below.)

table 15 of chapter 3

Tippett's Warp Breaks in Square-Roots

	BH	AM	AH	BM	BL	AL	(*)	(**)	(All)
(a) SQUARE-ROOTS OF COUNTS, ORDERED									
	5.3	6.00	6.6	6.5	6.6	8.4			
	4.9	5.9	6.0	6.2	6.4	8.2			
	4.6	5.5	5.3	6.2	5.7	7.3			
	4.5	5.4	5.1	5.4	5.4	7.2			
	4.1	4.6	4.9	5.3	5.4	7.1			
	4.0	4.2	4.6	5.1	5.2	5.5			
	3.9	4.2	4.2	4.6	4.5	5.1			
	3.9	4.1	3.9	4.4	4.4	5.1			
	3.6	3.5	3.2	4.0	3.7	5.0			
(b) VALUES NEEDED FOR SCHEMATIC SUMMARIES***									
end	5.3	6.0	6.6	6.5	6.6	8.4	(6.3)	(6.55)	
next	4.9	5.9	6.0	6.2	6.4	8.2	(5.95)	(6.3)	
quarter	4.75	5.7	5.65	6.2	6.05	7.75	(5.68)	(6.12)	
midmean***	4.22	4.77	4.83	5.32	5.26	6.46	(4.80)	(5.29)	(5.14)
quarter	3.9	4.15	4.05	4.5	4.45	5.1	(4.10)	(4.48)	
next	3.9	4.1	3.9	4.4	4.4	5.1	(4.0)	(4.4)	
end	3.6	3.5	3.2	4.0	3.7	5.0	(3.35)	(3.85)	
(c) SAME VALUES WITH MIDMEANS SUBTRACTED									
end	1.08	1.23	1.77	1.18	1.34	1.94	(1.50)	(1.26)	(1.42)
next	0.68	1.13	1.17	0.88	1.14	1.74	(1.15)	(1.01)	(1.12)
quarter	0.53	0.93	0.82	0.88	0.79	1.29	(0.88)	(0.83)	(0.88)
midmean***	0.00	0.00	0.00	0.00	0.00	0.00	(0.00)	(0.00)	(0.00)
quarter	-0.32	-0.62	-0.78	-0.82	-0.81	-1.36	(-0.70)	(-0.81)	(-0.78)
next	-0.32	-0.67	-0.93	-0.92	-0.86	-1.36	(-0.80)	(-0.89)	(-0.84)
end	-0.62	-1.27	-1.63	-1.32	-1.56	-1.46	(-1.45)	(-1.44)	(-1.31)

---

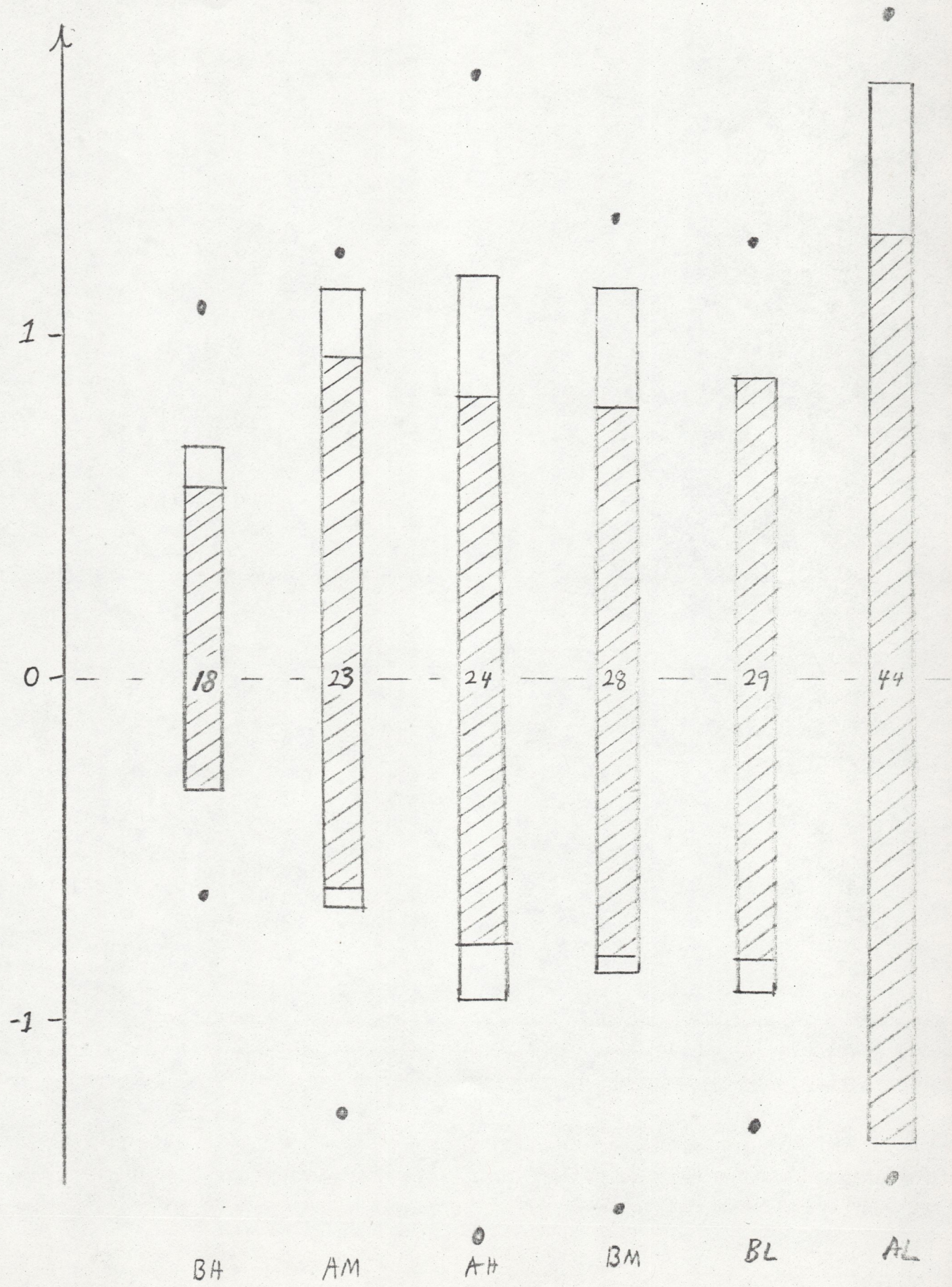
*Combination of values for AM and AH (for later use).

**Combination of values for BM and BL (for later use). Example: $6.12 = \frac{1}{2}[\frac{1}{2}(6.2) + \frac{1}{2}(6.2)] + \frac{1}{2}[\frac{1}{2}(6.4) + \frac{1}{2}(5.7)]$

***Exact midmean used.

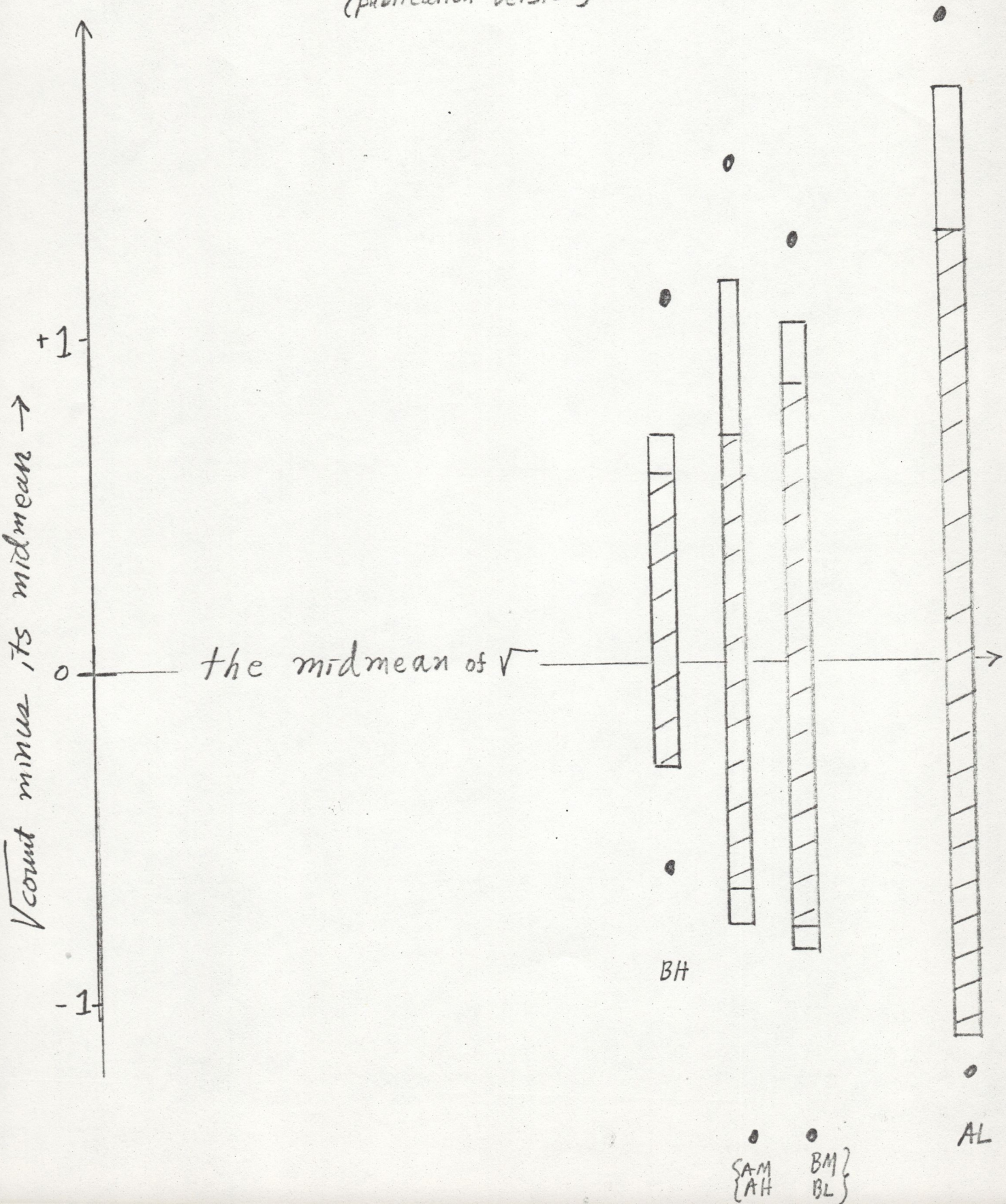
(All) = Mean of values in first six columns.

Tippett's warp breaks in square roots referred to midmeans



plot 17 of chapter 3

Tippett's warp breaks, in square roots, referred to -- and plotted against -- midmeans (of these square roots)
(Two single batches and two combined batches)
(publication version)



If we had had no signal from change of spread with change of midmean, we would have had to guide ourselves by the indications of symmetry and its lack. Table 18 sets out the numbers which it is most natural to examine when we are concerned with square roots: the sums of corresponding deviations from the midmean. We see a definite, but not impressive, tendency for these sums to be +, indicating a tendency for the upper parts of the batches to spread further than the lower parts. This, too, would call for moving further down the scale of powers, in the direction of using logs rather than square roots. (We shall still leave this exploration to the reader.)

3E. A rough look for stragglers

Our plots have, from time to time, caused us to suggest that certain values were "far out". To believe such a statement is not to say that these values are "wrong" or that they should necessarily be dropped or set aside. After all, one reason for using midmeans, next ranges and midspreads is to greatly reduce the need for explicit dropping or setting aside.

To believe that a value is "far out" is to believe that it deserves special attention and inquiry. Many "far out" values turn out to involve correctable errors of copying or arithmetic. Finding and eliminating these is desirable. Some "far out" values are signals of new phenomena or unsuspected conditions. Inquiries that show up their nature are very

table 18 of chapter 3

Indications of Asymmetry for Tippett's Data
in Square Roots

	Sums of deviations from midmeans						
	BH	AM	AH	BM	BL	AL	(*)
Ends	+.46	-.04	+.14	-.14	-.22	+.48	(+.09)
Nexts	+.36	+.46	+.24	-.04	+.28	+.38	(+.32)
Quarters	+.21	+.31	+.04	+.06	-.02	-.07	(+.06)

(*) Exact midmeans

Examples (see table 15(c) for starting values)

+.46 = (+1.08) + (-0.62)

+.36 = (+0.68) + (-0.32)

+.25 = (+0.57) + (-0.32)

rewarding. To be "far out" is to be a useful target for inquiry -- no more, no less. Not all "far out" values are "way out".

The identification of "far out" values is worthwhile enough for us to fix upon a definite way to do it. Looking at schematic summaries is easy if done crudely without need for agreement. It will continue to serve us. But a standard tool has its place also, saving thought and discussion for the inquiry into the "far out" values rather than wasting them on the question of which values are "far out".

We cannot expect any tool to draw a clear line between casual accident and important disparity. A dull tool -- used to warn us to look again and use whatever evidence and insight we can in thinking further -- meets our needs about as well as a tool can. Accordingly, we shall content ourselves with a simple way of looking for signs of a single "far out" value, or a few such -- a simple way of looking for values that, by comparison with the central cluster of values, appear to be so far away as to deserve attention.

If we turn back to plot 16 or plot 17, and ask ourselves why these symbolic summaries tended to make us think certain values were "far out", we are likely to find that what influenced us was the separation of the individual values from the filled-in part of the bar, judged in terms of the length of the bar.

This suggests the following simple process:

- find the quarter points or the hinges, then
- find the corresponding midspread, then

- lay off one midspread outward from each quarter point or hinge, then

- regard any value outside of the intervals thus covered as "far out".

This process seems to behave about as we might like. In many well-behaved situations it makes about one value in 20 "far out", which is about as many as we would like to see. If we flagged many fewer values our chance of making inquiries when they are badly needed would decrease, and would keep us less awake. If we flagged many more values the effort of inquiry would increase enough to make adequate inquiry almost intolerable, so that we would either suffer or fail to inquire deeply enough.

Let us go through the arithmetic, using data on the cost of plastering walls collected by the Bureau of Labor Statistics in 1923. Table 19 shows the result, which can be put in words thus:

- look at job 5 and ask if there is a reason for the low cost.

- look at job 11 and, especially, at job 2, and ask if there is a reason for the high cost.

Like most elementary procedures, we could probably have bypassed this one by staring hard enough at the numbers long enough. This routine is simple, however, so simple that it may be quicker than staring hard. Certainly it takes less thought.

Note that a man in a hurry would have used the hinges, here 0.27 and 0.29, in place of the quarter points, and found cutting points of 0.25 and 0.31. In this particular instance, the list of "far out" values would stay the same.

table 19 of chapter 3

Looking at 13 Costs of Plastering (Chicago, 1923)
in Search of Stragglers

(1) Job number.

(2) Cost of plastering (in 1923 dollars per square yard).

(3) Costs arranged in order; the ← mark the 14 (= 1+13) intervals defined by the 13 costs.

(4) Quarter points, .270 and .295; the midspread intervals (see text), and difference between these quarter points (in parenthesis), (.025) = .295 - .270; one midspread further out each way to the cutting points, also in parenthesis, .270 - .025 = (.245), .295 + .025 = (.320). The ! mark three possible stragglers that deserve being thought about.

(1)	(2)	(3)	(4)
		←	
1	.28	.23	!
		←	(.245) cutting point
2	.49	.27	
		←	(.025)
3	.29	.27	
		←	$.27 = \frac{.27+.27}{2}$ quarter point
4	.28	.27	
		←	
5	.23	.28	
		←	
6	.28	.28	
		←	(.025)
7	.27	.28	
		←	
8	.28	.28	
		←	
9	.30	.29	
		←	
10	.27	.29	
		←	$.295 = \frac{.29+.30}{2}$ quarter point
11	.38	.30	(.025)
		←	(.320) cutting point
12	.27	.38	!
		←	
13	.29	.49	!
		←	

From (?) the Monthly Labor Review for November 1924. (They also appear at pages 53-62 of F. C. Mills and D. H. Davenport, A Manual of Problems and Tables in Statistics with Notes on Statistical Procedure (New York, 1925, Henry Holt).

Using this tool is not to be the "end of the line". If we find an "apparent straggler" we are to look at it, think about it, perhaps conclude that it seems to be either meaningfully extreme or accidentally extreme. Such conclusions are to come from other evidence and insight, never only from the fact that it seemed to stick out. To catch the smaller meaningful deviations, we have to catch many intermediate-sized accidental deviations. We may have to depend on follow-up "field" studies to indicate which are which.

Sometimes differences between adjacent observations tend to be narrower toward one end of the batch and broader toward the other. This is not easy to detect. Usually either many batches or a very large batch is required to do very well at detection. If the observations called to our notice are repeatedly on one end rather than -- or more frequently than -- on the other, it often pays to ask whether we cannot gain by reexpressing the observations so that the numbers describing them are more symmetrically patterned. Improving the way observations are expressed as numbers is one of the important things we can do.

3F. Typical squares and mean squares

The indications of spread we have so far set up are enough to serve our present simple needs. As we go on to later chapters, however, there will be new needs and we will find still other measures of spread convenient. In part some of this convenience is purely historical. Some it is purely

convenience in algebraic manipulations and handy identities. Some may run deeper than any of these. Whatever the real reasons, we will want to use these additional measures later, so we shall explain and illustrate them now.

The first of these, the <u>typical square</u>, is merely the sum of the squares divided by the number of entries. For residuals from 19, Tippett's nine counts of warp breaks under condition BH give:

$$\frac{(+9)^2 + (+5)^2 + (+2)^2 + (+1)^2 + (-2)^2 + (-3)^2 + (-4)^2 + (-4)^2 + (-6)^2}{9}$$

$$= \frac{192}{9} = 21.3$$

The interpretation of the typical square is simple. 21.9 is the square of ±4.68; its appearance as the typical square indicates that the sum of squares of these residuals is the same as if every one of them were -4.68 or +4.68.

We can apply the same idea to the nine original counts, finding

$$\frac{28^2+24^2+21^2+20^2+17^2+16^2+15^2+15^2+13^2}{9} = \frac{3365}{9} = 373.9$$

as the typical square of the original counts. The typical square of 19 is clearly 361. Thus we have the following typical squares:

original counts 373.9

and

residuals from 19	21.3
19 as typical value	361.
	385.3

We see that the sum of the typical squares for the pieces:

- typical value,
- residuals,

adds up to something close to the typical square for the original counts. If we had used the mean of the original counts, 18.78, as our typical value -- both to be squared and to define our residuals -- the sum of the typical squares for the pieces would have equaled the typical square for the original counts. If we were doing hand arithmetic, such "adding up exactly" behavior would offer us our choice of two computationally useful things:

- not having to calculate residuals from the exact mean in order to find their typical square, since we could find this typical square by difference, <u>or</u>
- having a check on our arithmetic.

As machine computation takes over more and more of the load, both these considerations matter less, and our interest in exactness of adding up fades away. Even when we use 19 as a basis for finding residuals, the component typical squares:

361	for typical
21	for residuals

give us a useful picture of the relative sizes of the two components.

(Sometimes it will be important to take the square roots and look at 19 and 4.6. These square roots are on the scale of the counts themselves, but we cannot expect them to do anything like adding up. (In the example, we have $\sqrt{373.9} = 19.3$ vs $19 + 4.6 = 23.6$.)

In addition to typical squares, there are also <u>mean squares</u>, in which we divide by a number that is usually smaller than the total number of items. The rule for finding this denominator, often rather mysteriously called "degrees of freedom", can be expressed, more or less precisely, in several ways, including:

- take the number of values and subtract the number of constants already fitted and subtracted (works in reasonable generality).
- take the number of values that can be freely set at desired values (works best when means and other similarly behaving indicators are used to define residuals).

Whatever words we use, the divisors and mean squares corresponding to our three examples will run as follows:

For	Divisor*	Mean square
Original counts	9	$\frac{3365}{9} = 374$
Residuals	9-1 = 8	$\frac{192}{8} = 24$
Typical	1	$\frac{3249}{1} = 3249$

*Called degrees of freedom.

Notice that in this example, as in many others, the degrees of freedom really do add up.

An adequate explanation of why mean squares can aid data analysis takes more background than we have now -- or will get in this account. Some of the concepts on which confirmatory data analysis is founded are required. Accordingly, we shall not try to give an explanation, limiting ourselves to a hint.

Suppose we have a batch of n observations, all but one equal to zero. (This is about as irregular and unsystematic behavior as we can get.) If the values are 0, 0, ..., 0, c, then the mean and the residuals from the mean behave as follows:

Component	Typical square	Mean square
Mean $\left(= \frac{c}{n}\right)$	$\frac{c^2}{n^2}$	$\frac{c^2}{n}$
Residuals from mean	$\frac{(n-1)c^2}{n^2}$	$\frac{c^2}{n}$
(Original values)	$\left(\frac{c^2}{n}\right)$	$\left(\frac{c^2}{n}\right)$

Notice that the two typical squares do add up to c^2/n, the typical square for the original values, while the two mean squares mark the utter irregularity of the observations by coming out equal, both to each other and to the mean square for the original value.

Both typical squares and mean squares look hard at every value, even the more extreme ones. When we are dealing with values centered around zero, there is often a need for a related indicator of size which avoids the most extreme values. The simplest solution is to use

$$\text{midsquare} = \tfrac{1}{2}(\text{midspread})^2$$

which is often close to the typical square in size when the batch shows no distinctive appearance. For the bowling scores referred to their mean of 198, namely -14, -13, -7, -6, -5, 0, 5, 9, 12, 16 we find (to the nearest whole number):

$$\text{typical square} = 98$$

$$\text{easy midsquare} = \tfrac{1}{2}(16)^2 = 128$$

$$\text{exact midsquare} = \tfrac{1}{2}(18.25)^2 = 167$$

Clearly this batch of 100 bowling scores shows no tendency for the mean square to far exceed the midsquare. As we were already convinced, there is no indication of straggling among the 10 bowling scores.

3G. <u>How far have we come</u>?

This chapter has introduced us to:

- a variety of measures of spread.
- plotting batches for comparison and examination.
- the effects of choice of expression and how such choices can help.
- how to identify some "far out" values easily.

At first glance these may seem a little diverse, but only a moment is needed to see their connections. Without measures of spread, we could not make the plots we use to represent batches, we could not see important effects of choices of expression, we could not identify "far out" values easily.

The chapter could reasonably have been labeled "Measures of spread, and some of their simple consequences".

The range, next range, easy midspread and exact midspread all have reasonably simple definitions. One or another of them will meet almost any need we meet in exploring simply structured data. The elementary but important things already said about them should not need resaying here.

The simple summary scheme we use for portraying batches makes use of the end values, next-to-end values, hinges, and quarter points from which these measures of spread are calculated, and the midmean to reveal much about the batch (little about any tendencies to separation into sub-batches). The more we use them, the more they will tell us.

In chapter 2 we sought straightness and flatness, as we still sometimes do in this chapter. Here, however, we more typically seek spreads which don't drift systematically as the midmean changes, or even -- when nothing better can be found -- symmetry of pattern.

The problem of which values are "way out" in some kind of erratic, or wild way, and which are "far out" because of truly accidental circumstances is an old one. We, like all our predecessors, have no complete solution. We have suggested how to label some values as "far out" by comparison with the hinges or quarter points. Such values are not thereby labeled "way out". Rather they are marked for special attention -- said to be worthy of special attention and consideration, especially when this attention gives us ground for judging whether each individual is "way out" or "accidentally extreme".

Finally, we have introduced a group of spread-like measures of size: mean squares, typical squares, and midsquares, whose usefulness will slowly become more and more apparent as we proceed.

3K. How to choose the next trial expression using spread

In trying to straighten and flatten plots of relationship, we can make rather direct use of various devices as we saw in chapter 2. In this chapter, most reformation of expression has been based on a desire to make spreads roughly equal, or -- if no better guide can be found -- to promote symmetry in the extent of the two wings. How do we use unsatisfactory appearances -- spread apparently changing with midmean, or apparent lack of symmetry -- to decide which expression to try next? To expect a simple and complete answer to this question is to ask too much. To hope for some guidance is to be reasonable and wise.

Let us recall plot 17, in which schematic summaries of spread were plotted against the value of the midmean. We saw more or less of a tendency to radiate from the origin, and decided to take logs. Was this reasonable? Suppose spreads are uniform in logs, how would they appear in raw values? Since constancy in difference of logs is the same as constancy in %, the following exhibit equal spread in logs:

8, 9, 10, 11, 12

16, 18, 20, 22, 24

24, 27, 30, 33, 36

32, 36, 40, 44, 48

40, 45, 50, 55, 60

If we treat the values in the middle column as the midmeans, subtracting them from the other values in the line and also using them (along the horizontal scale) for plotting, we find the result in the upper panel of plot 20.

If spread were ideally constant in square roots, we might see values like these:

64,	81,	100,	121,	144
144,	169,	196,	225,	256
225,	256,	289,	324,	361
324,	361,	400,	441,	484
400,	441,	484,	529,	576

Plotting deviations against midmeans gives, with appropriate changes of scale the picture in the middle panel in plot 20.

If spread were constant in terms of the values we are presently working with -- reciprocal square roots -- we would find a picture like that in the lowest panel of plot 20.

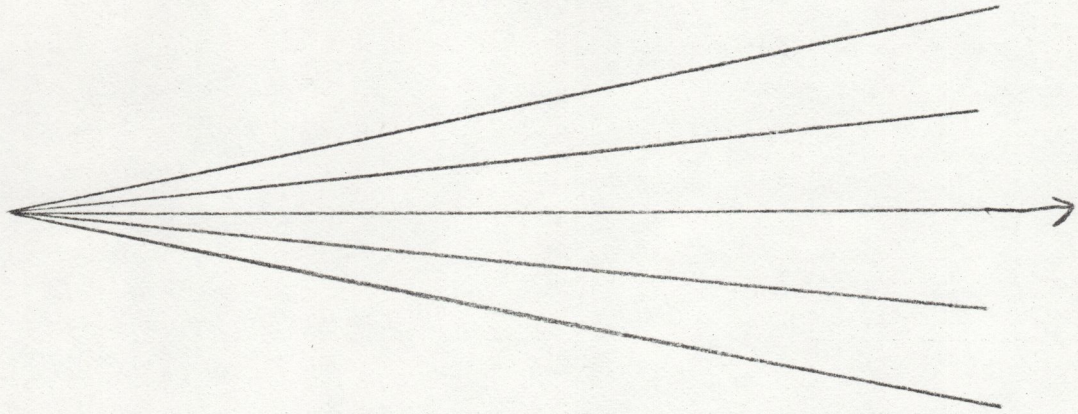
The situation is easy to summarize:

- if logs are needed, the ideal picture is triangular.
- if square roots are needed, the ideal picture has a blunt nose.
- if some negative power is needed, the ideal picture is sharper nosed than a triangle.

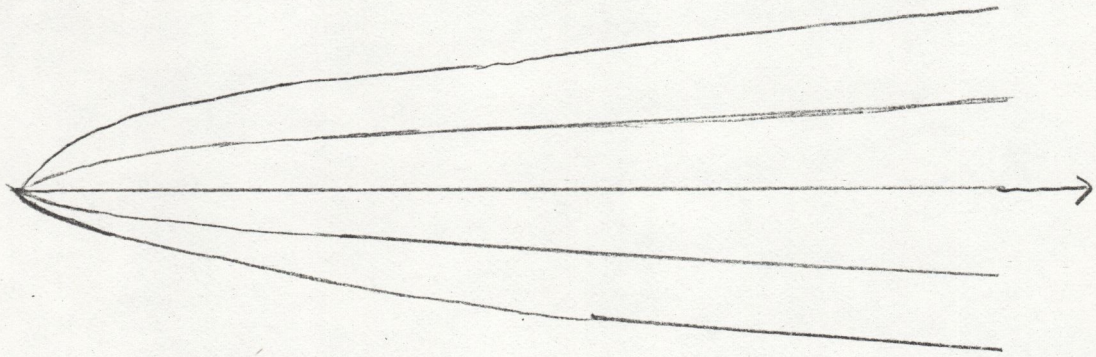
If we have enough data, behaving consistently enough, we have enough guidance from such a deviation-vs-midmean plot to do fairly well in choosing the next mode of expression to try.

plot 20 of chapter 3

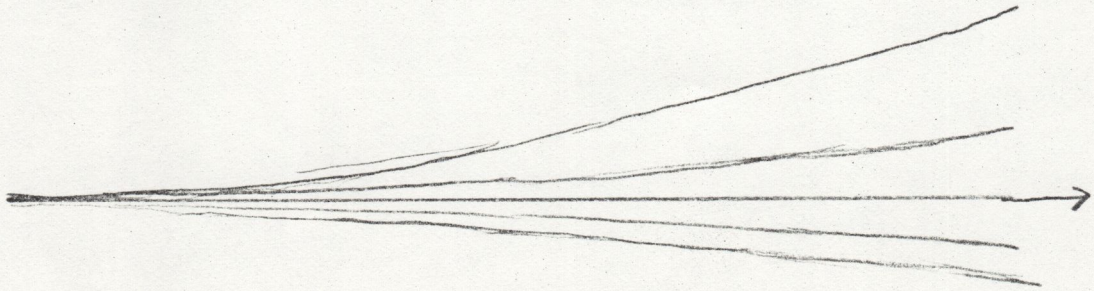
Behavior of deviations plotted against midmeans (on uniform scales) when spreads are constant for each of various other expressions of the data. ("the data" being whatever expression was used to find midmeans and then deviations. (Ideal instances.)



Spread constant in logs of data



Spread constant in $\sqrt{\ }$'s of data



Spread constant in $-1/\sqrt{\ }$'s of data

Turning back to Tippett's warp breaks and plot 17, we now see that the triangular appearance which appeared to be present was a reasonable indication that we ought to try using the log of whatever expression had been used to calculate its deviations and midmeans. Since we were using

$$\sqrt{\text{count}}$$

this tells us to use

$$\log \sqrt{\text{count}} = \tfrac{1}{2} \log \text{count}$$

where the factor $\frac{1}{2}$ can be neglected as a scale factor. Thus the message of the appearance of plot 17 was "use logs". In making this translation, however, we need to notice that the ratio of largest midmean to smallest midmean is only about 3 to 2 in plot 17, so that we cannot be too sure that logs of what plot 17 is based on is the right choice -- square roots of what plot 17 is based on would not be too bad an account of plot 17's appearance. Thus

$$\text{square root of } \sqrt{\text{count}} = \sqrt[4]{\text{count}}$$

is by no means out of possibility for the expression that would make spread constant.

3L. How to choose the next trial expression using symmetry (if you must)

What if we must be guided by symmetry? Plotting techniques do not seem essential, and straight numerical trial is relatively easy. If we try logs, for example, we have

$$\text{log end value} \equiv \text{end log value}$$

although

$$\text{log mean of values} > \text{mean of log values}$$

(unless all values are equal) as when

$$\frac{1+5}{2} = 3, \quad \log 3 = .47712$$

but

$$\frac{\log 1 + \log 5}{2} = \frac{.69897}{2} = .34948$$

illustrates.

End values, and next-to-end values behave simply when we take logs -- or square roots -- or what have you. Means and midmeans do not, nor do (when they have to be interpolated) hinges and quarter points. However, with the great exception of the mean, the differences between, for example, log quarter points and quarter points of log values, are usually small enough to make the easy calculation quite feasible as a working approximation.

Table **21** illustrates the calculations for (a) Tippett's BH batch and (b) the mean of batches AM and AH. Instead of calculating and comparing deviations, this table takes the arithmetically simpler rate of summing each pair of hopefully symmetrical values and comparing the result with twice the midmean. We have been careful to use the minus signs on the reciprocal expressions, so that the direction of tilt does not change secretly adjacent to the logarithm. In panel (a),

table **21** of chapter 3

Exploration of Reexpression to Gain Symmetry

(1) Values based on: Square roots of counts, as in table 15(a); our data

(2) Values based on: Square roots of data

(3) Values based on: Log of data

(4) Values based on: Reciprocals of square roots of data

(5) Values based on: Reciprocals of data

	(1) $\sqrt{\cdot}$	(2) $\sqrt{\sqrt{\cdot}}$	(3) $\log\sqrt{\cdot}$	(4) $-1/\sqrt{\sqrt{\cdot}}$	(5) $-1/\sqrt{\cdot}$
(a) BATCH BH ALONE					
end	5.3	2.30	.724	-.435	-.189
next	4.9	2.21	.690	-.452	-.204
quarter	4.75	2.18	.677	-.459	-.211
midmean	4.22	2.05	.625	-.488	-.237
quarter	3.9	1.97	.591	-.508	-.256
next	3.9	1.97	.591	-.508	-.256
end	3.6	1.90	.556	-.526	-.278
sum of ends	8.9	4.20	1.280	-.961	-.467
sum of nexts	8.8	4.18	1.281	-.960	-.460
sum of quarters	8.65	4.15	1.268	-.967	-.467
2 x midmean	8.44	4.10	1.250	-.976	-.474
(b) BATCHES AM AND AH COMBINED					
end	6.3	2.51	.799	-.398	-.159
next	5.95	2.44	.775	-.410	-.168
quarter	5.68	2.38	.754	-.420	-.176
midmean	4.80	2.19	.681	-.457	-.208
quarter	4.1	2.02	.613	-.495	-.244
next	4.0	2.00	.602	-.500	-.250
end	3.35	1.83	.525	-.546	-.299
sum of ends	9.65	4.34	1.324	-.944	-.458
sum of nexts	9.95	4.44	1.377	-.910	-.418
sum of quarters	9.78	4.40	1.367	-.915	-.420
sum of midmeans	9.60	4.38	1.362	-.914	-.416

the sums (of ends, nexts, and quarters) are higher than twice the midmean for raw, $\sqrt{\ }$, and log expressions. They are lower for reciprocal expression. We can only say that panel (a) points -- certainly not very definitely -- toward a reciprocal square root. For panel (b), twice the midmean matches the ends for an expression intermediate between raw and square root. It matches the nexts between the reciprocal square root and the reciprocal. It matches the quarters between the logarithm and the reciprocal square root. Just where panel (b) points is hard to say.

We need to emphasize that these calculations are approximate and at best suggestive. If we make the reexpressions we have been considering, the midmeans, for example, will come out slightly differently. There is thus no substitute for a complete reexpression, with recalculation, once we are persuaded which one to try next.

Especially in view of this, almost anything we can do to speed up and simplify our explorations is worthwhile. In particular, we do not want to do either six or even four panels' worth of computation on Tippett's data if we can get along with less. One way to get along with less is to make up a composite set of values as follows:

- form deviations from midmeans (done in table 14(15))
- form the mean across batches (also done there)
- find mean of midmeans (also done there)
- restore mean of midmeans to mean of deviations to provide composite values.

Table 22 illustrates this assembly process, followed by exploration of the same set of possible reexplorations.

table 22 of chapter 3

Exploration of Reexpression for a Composite
(based on table 14(15))

(1) Raw values
(2) Square roots of data
(3) Logs of data
(4) Reciprocals of square roots of data
(5) Reciprocals of data

	(1)	(2)	(3)	(4)	(5)
(a) MEAN OF SIX DEVIATIONS FROM MIDMEANS "(All)" FROM TABLE 15					
end	1.42				
next	1.12				
quarter	0.88				
midmean	0.00				
quarter	-0.78				
next	-0.84				
end	-1.31				
(b) SAME WITH MEAN OF MIDMEANS, 5.14, RESTORED					
end	6.56	2.56	.817	.390	.152
next	6.26	2.50	.797	.400	.160
quarter	6.02	2.45	.780	.408	.166
midmean	5.14	2.27	.711	.441	.195
quarter	4.36	2.09	.640	.479	.229
next	4.30	2.07	.634	.482	.233
end	3.83	1.93	.571	.518	.268
sum of ends	10.39	4.49	1.388	-.908	-.420
sum of nexts	10.56	4.57	1.431	-.882	-.393
sum of quarters	10.38	4.54	1.420	-.887	-.395
2 x midmean	10.28	4.54	1.422	-.882	-.390

Note: Entries in (a) and 5.14 are from the "(All)" column of table 14(15).

Examples: 6.56 = 1.42 + 5.14, 3.83 = -1.31 + 5.14.

Again, ends, nexts, and quarters match the doubled midmean for somewhat different choices, with the square root seeming a reasonable compromise.

The weakness of the evidence offered by symmetry and its lack is clear in this example. We should:

- only use the approach of this section when no better basis for choosing an expression is available.
- always plan to check up by making the next trial, and looking to see how symmetrical its results are.
- always treating the expression indicated by symmetry or its lack as rather tentative, unless it is based on an unusually large collection of data.

3X. Another example

In 1950, Bruner, Postman, and Mosteller reported on an experiment to study and clarify a psychological phenomenon, the reversal of perspective. Certain pictures can be seen as in either of two perspectives; most subjects see first one, then the other, so that it is natural to measure how rapidly the perspective reverses. In the experiment this was done separately for each minute of a 10-minute trial, 19 subjects being studied under each of three sets of instructions:

- "alternate"
- "hold"
- "natural"

We shall begin with the last eight (of the 10) minutes under the "alternate" instruction, looking at the even-numbered subjects only. (The first two minutes are omitted because they may show "start-up" phenomena.) Plot 23 shows the schematic summaries.

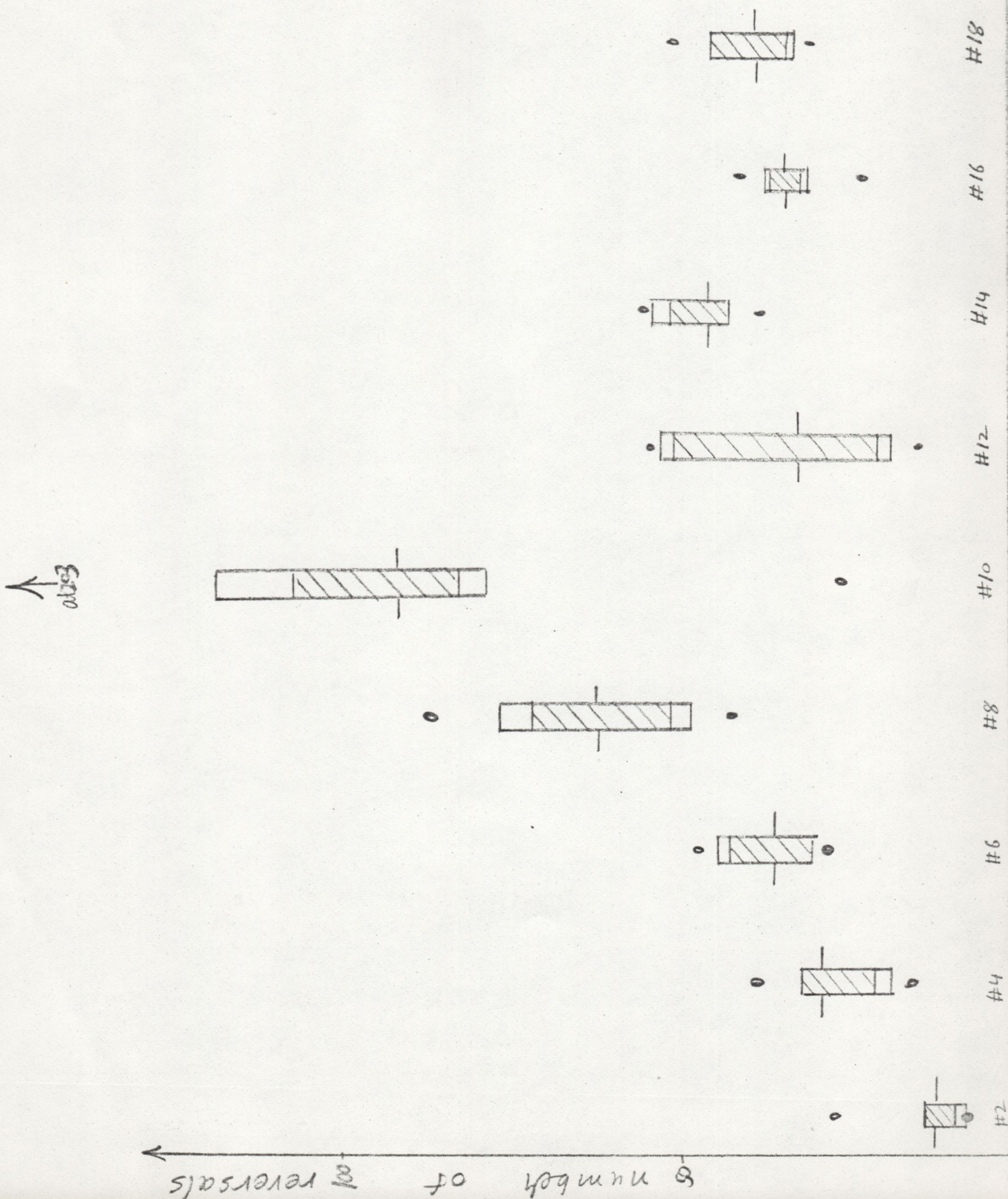
One of the first things that strikes us is the combination of a long body with low values for subject 12. Turning back to the original data, we find that his ten counts, in order of occurrence were:

30, 22, 14, 22, 24, 32, 18, 48, 52, 53

It would appear that after seven minutes this subject suddenly learned how to make his perspective reverse rapidly. We shall have to allow for this as we proceed.

plot 23 of chapter 3

Schematic summaries for 9 subjects when instructed to "alternate".



Since the typical number of alternations varies widely, we naturally wish to see whether other characteristics of the counts vary with it. Plot 24 shows the same displays for all even-numbered subjects except #12, with the subjects ordered according to midmean. This plot strongly suggests that:

- larger spreads appear with larger midmeans.
- the lowest and highest values for subject 10 are anomalous, even when we allow for the large midmean.

To look more closely at the spreads of the various individuals, we turn to plot 25, which plots midspread against midmean for each batch of eight counts. The suggestion of a relationship appears strengthened.

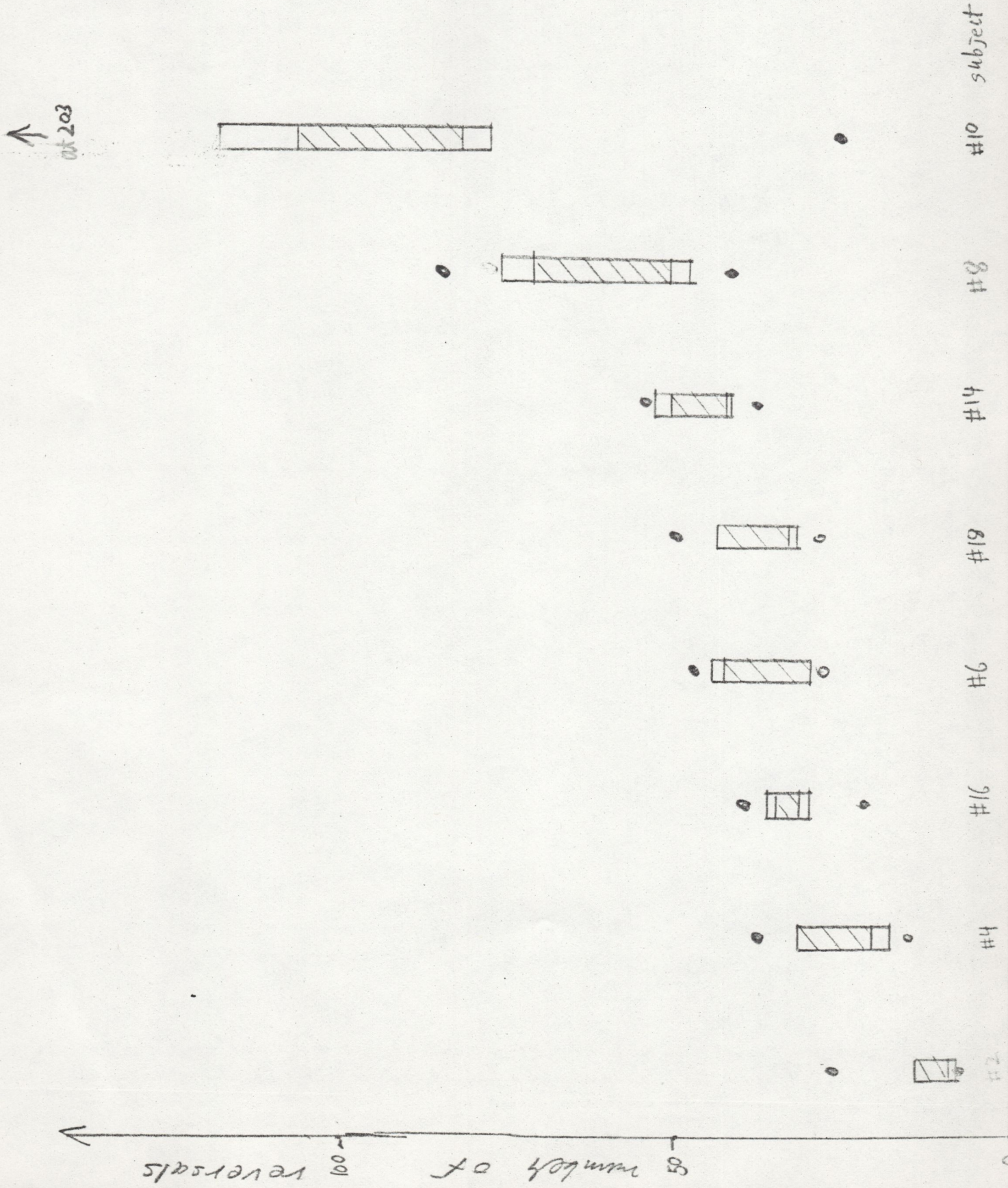
The natural ways to reexpress counts are, we believe, square roots or logs. Plot 26 and plot 27 show the midspreads for these expressions of the data. In these plots, in order to simplify and clarify comparing one plot with another, we have used the midmeans of the rows as the horizontal scale. We see that:

- the spreads of the square roots still seem to trend upward.
- the spreads of the logs seem to trend downward if they trend at all.

Some change in expression is indicated, though the amount of data we have been exploring is not enough to give a clear indication. Logs look a little preferred as compared with square roots, but square roots might well prove quite acceptable.

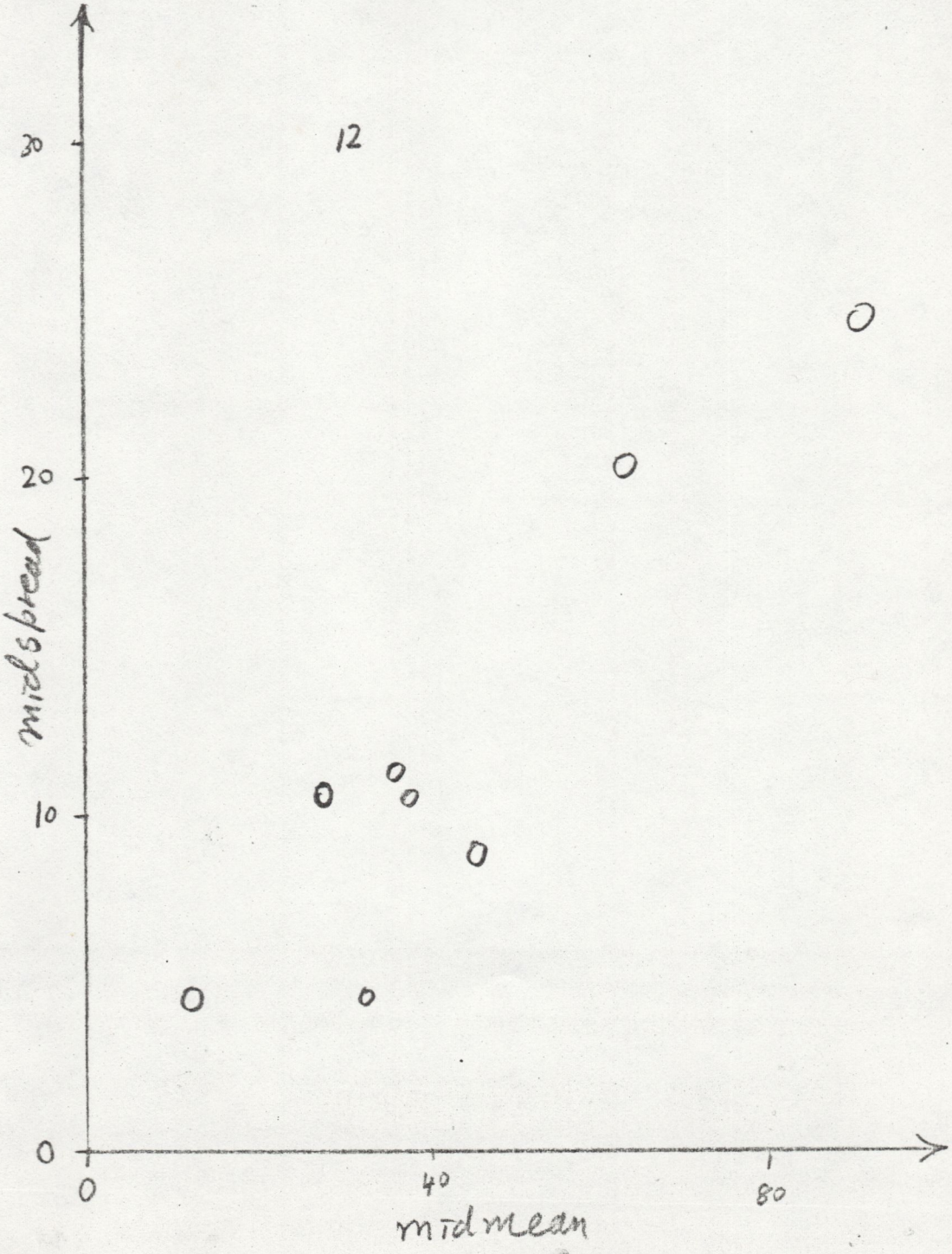
plot 24 of chapter 3

Schematic summaries ordered by midmean for 8 of the same subjects

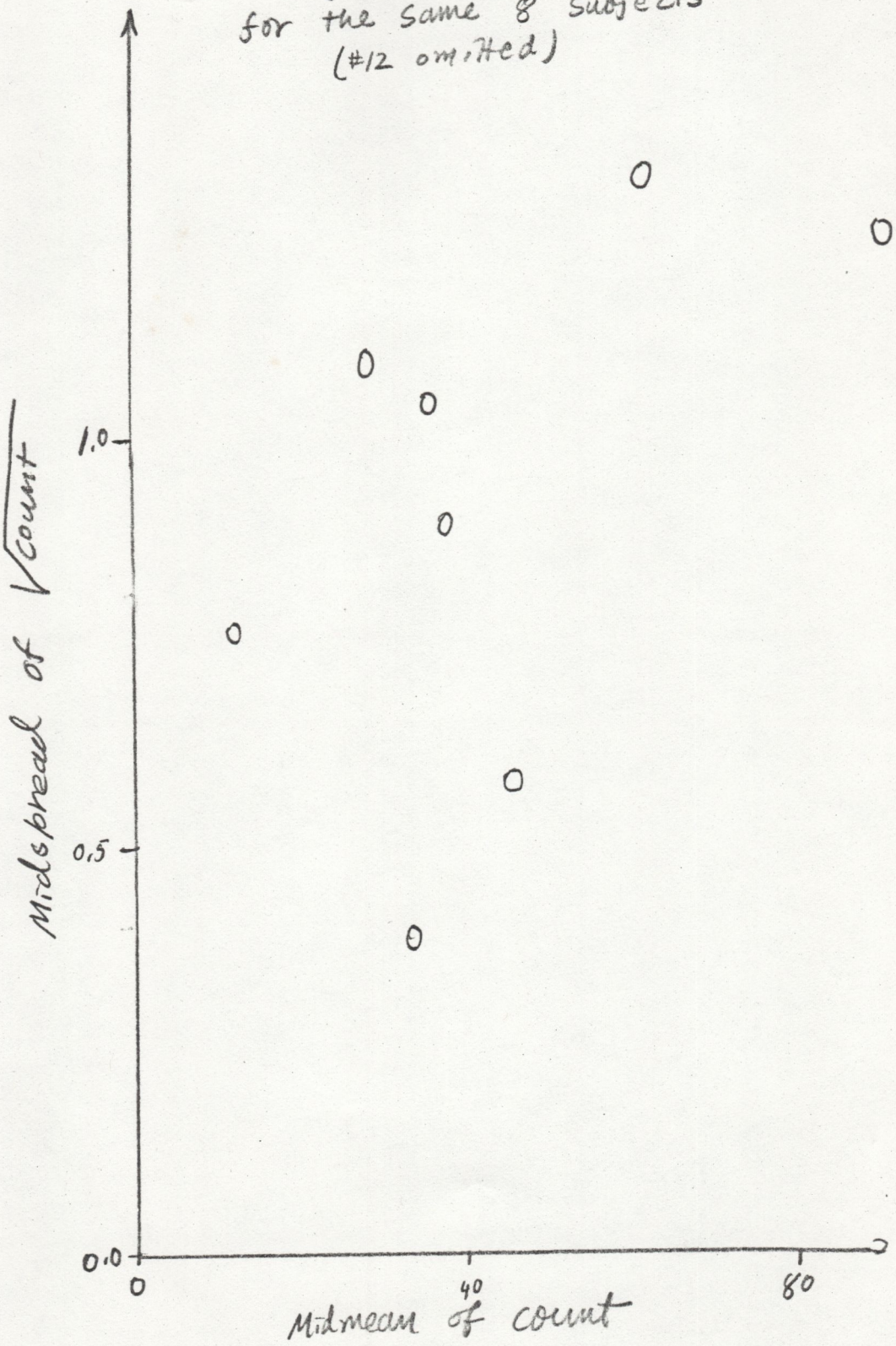


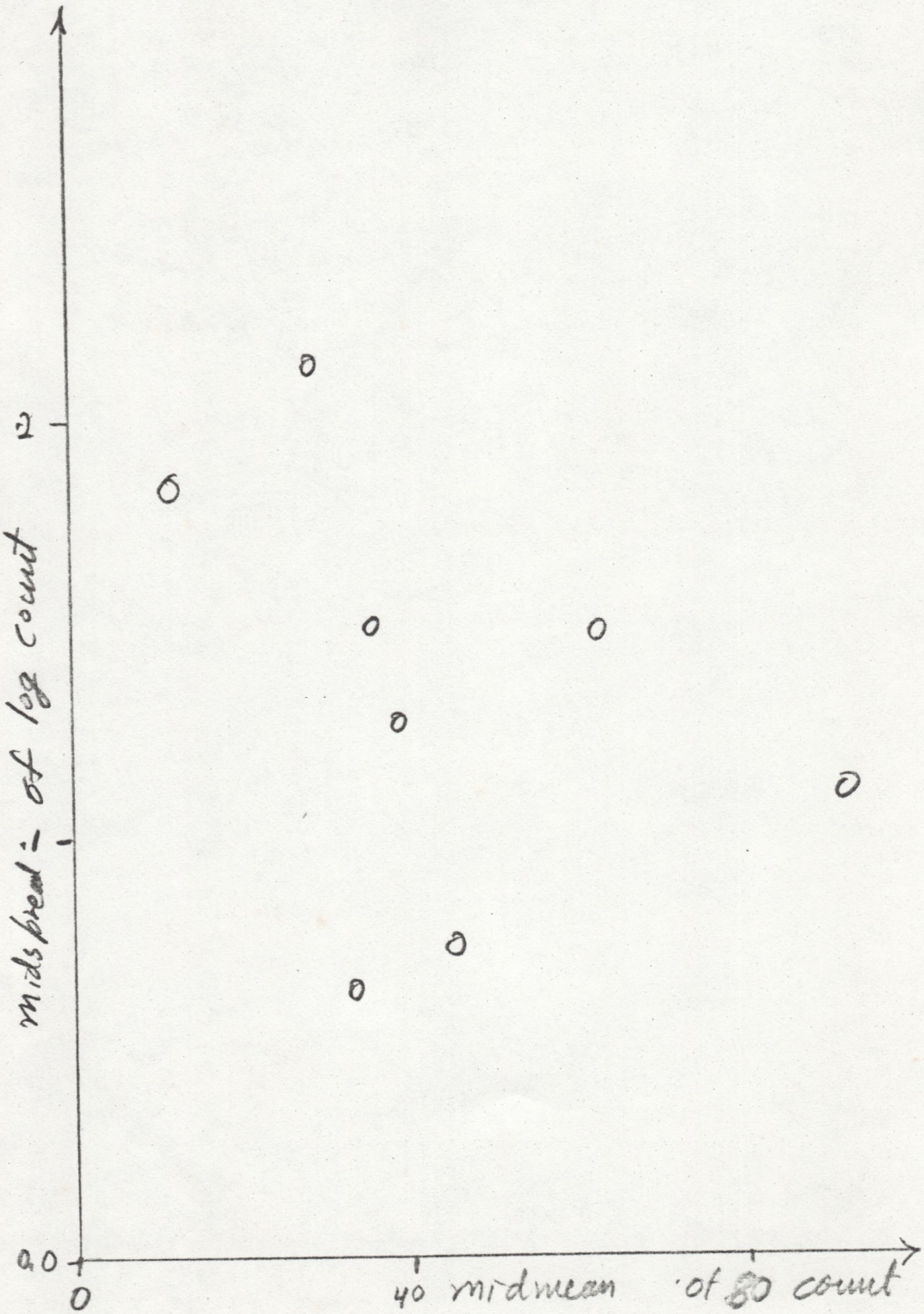
plot 25 of chapter 3

Midspread plotted against midmean
(for all 9 subjects, #12 plotted as "12")



Midspread of $\sqrt{\text{count}}$ against midmean (midmean of count) for the same 8 subjects (#12 omitted)



plot 27 of chapter 3

Midspread against midmean for the 8 subjects (omitting #12). (Midspread of log counts)